

AD-A145 545

IDENTIFICATION AS A FUNCTION OF FAMILIARITY FOR KNOWN
VOICES TALKING OVER..(U) NAVAL RESEARCH LAB WASHINGTON
DC A SCHMIDT-NIELSEN ET AL. 16 JUL 84 NRL-MR-5382

1/1

UNCLASSIFIED

SBI-AD-E000 591

F/G 5/8

NL

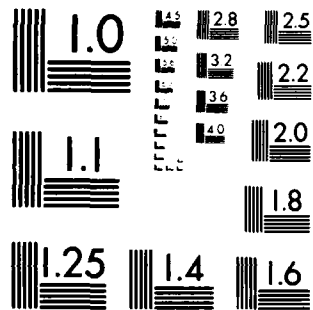
END

DATE

FILED

10-84

DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

AD-A145 545

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE				
1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED		1b. RESTRICTIVE MARKINGS		
2a. SECURITY CLASSIFICATION AUTHORITY		3. DISTRIBUTION/AVAILABILITY OF REPORT		
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE		Approved for public release; distribution unlimited.		
4. PERFORMING ORGANIZATION REPORT NUMBER(S) NRL Memorandum Report 5382		5. MONITORING ORGANIZATION REPORT NUMBER(S)		
6a. NAME OF PERFORMING ORGANIZATION Naval Research Laboratory	6b. OFFICE SYMBOL (If applicable) Code 7520	7a. NAME OF MONITORING ORGANIZATION		
6c. ADDRESS (City, State and ZIP Code) Washington, DC 20375		7b. ADDRESS (City, State and ZIP Code)		
8a. NAME OF FUNDING/SPONSORING ORGANIZATION	8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER		
8c. ADDRESS (City, State and ZIP Code)		10. SOURCE OF FUNDING NOS.		
		PROGRAM ELEMENT NO 208010N	PROJECT NO X0919	TASK NO DN260-149
11. TITLE (Include Security Classification) (See page ii)				
12. PERSONAL AUTHOR(S) Schmidt-Nielsen, A. and Stern, K.R.				
13a. TYPE OF REPORT Interim	13b. TIME COVERED FROM 10/82 TO 3/84	14. DATE OF REPORT (Yr., Mo., Day) July 16, 1984	15. PAGE COUNT 37	
16. SUPPLEMENTARY NOTATION				
17. COSATI CODES		18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)		
FIELD	GROUP	SUB. GR.		
			Speaker recognition Linear Predictive Coding (LPC)	
			Human listeners Voice communications	
19. ABSTRACT (Continue on reverse if necessary and identify by block number)				
<p>A commonly cited drawback of narrowband systems such as the DoD standard linear predictive coding (LPC) algorithm is that speaker recognition is poor. Yet it is the opinion of many users that they frequently recognize the speaker. Tape recordings of 24 speakers conversing over an unprocessed channel and over an LPC voice processing system were subjected to listening tests. Twenty-four co-workers listened to the tapes and attempted to identify each speaker from a list of about 40 people in the same branch. Prior to the recognition tests, each of the listeners also rated his or her familiarity with each of the speakers and the distinctiveness of each speaker's voice. There was some loss in voice recognition over LPC, but the recognition rate was still quite high. Unprocessed voices were correctly identified 88% of the time, whereas the same people talking over the LPC system were correctly identified 69% of the time. Talker familiarity was significantly correlated with correct identifications. There was no significant correlation between the rated distinctiveness of the speaker and correct identifications. However, familiarity and distinctiveness ratings were highly correlated.</p>				
20. DISTRIBUTION AVAILABILITY OF ABSTRACT		21. ABSTRACT SECURITY CLASSIFICATION		
UNCLASSIFIED/UNLIMITED <input checked="" type="checkbox"/> SAME AS RPT <input type="checkbox"/> DTIC USERS <input type="checkbox"/>		UNCLASSIFIED		
22a. NAME OF RESPONSIBLE INDIVIDUAL A. Schmidt-Nielsen		22b. TELEPHONE NUMBER (Include Area Code) (202) 767-2682	22c. OFFICE SYMBOL Code 7520	

DD FORM 1473, 83 APR

EDITION OF 1 JAN 73 IS OBSOLETE

SECURITY CLASSIFICATION OF THIS PAGE

SECURITY CLASSIFICATION OF THIS PAGE

11. TITLE (Include Security Classification)

Identification as a Function of Familiarity for Known Voices Talking Over an Unprocessed Channel and an LPC Voice Processor

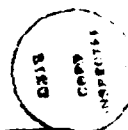
SECURITY CLASSIFICATION OF THIS PAGE

CONTENTS

INTRODUCTION	1
EXPERIMENT I	6
EXPERIMENT II	22
SUMMARY AND CONCLUSIONS	27
REFERENCES	30

S DTIC
ELECTE
SEP 4 1984
B **D**

Accession For	
NTIS GRA&I	<input checked="" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	



**IDENTIFICATION AS A FUNCTION OF FAMILIARITY FOR
KNOWN VOICES TALKING OVER AN UNPROCESSED
CHANNEL AND AN LPC VOICE PROCESSOR**

INTRODUCTION

In identifying an unseen speaker, people rely on a variety of voice characteristics and speech habits. A good voice communication system should not only provide good intelligibility, but should also preserve the identifying characteristics of each voice. The telephone is the most familiar and commonly used voice link, and most people have very little trouble recognizing familiar speakers over a normal phone connection. Even though telephone communication is generally considered to be of reasonably high quality, there is still some loss in voice recognition, as anyone who has failed to recognize the "telephone voice" of a new acquaintance can verify. Greater losses in quality can be expected in contexts (including many military applications) where the bandwidth is limited and a high data rate cannot be supported. The 2.4 kilobits per second (kbps) linear predictive coding (LPC) algorithm adopted as the DoD standard (Tremain, 1982) produces speech with good intelligibility; the score on the Diagnostic Rhyme Test (DRT) is 87.9% (Sandy, 1982). It is generally assumed that although the analysis and resynthesis process used in the LPC algorithm preserves the intelligibility of the speech sounds reasonably well, speaker recognition is severely impaired. Countering this is the anecdotal evidence of many users of LPC systems who claim that they do recognize individual voices. The purpose of this paper is to explore some aspects of the recognition of familiar speakers and to compare recognition over a voice system using a 2.4 kbps LPC algorithm with recognition of high quality tape recordings of the same speakers. The materials used in these tests were recordings of people talking to one another over a voice link in order to perform a communication task. The conversations were spontaneous in that they did not involve a written text.

Manuscript approved May 7, 1984.

In voice recognition the distinction has been made between recognition by machine, i.e. computer, (see Atal, 1976; Rosenberg, 1976 for a review) and recognition by listening (see Bricker and Pruzansky, 1975; Hecker, 1971 for a review). An alternative classification has been suggested by Nolan (1983, p. 7). Because the behavior of naive listeners differs drastically from the technical classifications that can be made aurally by expert phoneticians or visually by trained spectrogram readers, he suggests that technical speaker recognition and naive speaker recognition is a more satisfactory classification. Computer (or technical) voice recognition is usually divided into two categories: identification and verification. Technical identification can be used in criminal cases, while verification procedures are important for security purposes, such as access to bank accounts or privileged information. The same distinction may be applied to recognition by human listeners: do I know who this person is (identification), and does this sound like George (verification)? Both of these aspects are important for good acceptability of a voice system. In the course of everyday activities, the most common problem facing the human listener is that of identifying, or naming, the person whose voice is heard. The present research is concerned primarily with naive speaker recognition, and the technical aspects of verification (especially those accomplished more reliably by an automatic system) are outside the scope of this paper.

The voice characteristics that allow listeners to identify speakers have been divided into organic, those related to the anatomy of the vocal tract, and acquired, those that are learned over the course of one's life (Garvin and Ladefoged, 1963). Acquired voice characteristics are speech habits and

include such things as regional accent, vocabulary choice, interjections, hesitation patterns, etc. These are qualities that a mimic might imitate to create a plausible caricature of a person's voice. They are strong cues that are used often in everyday life for speaker recognition. Indeed, these cues are so obvious that in most voice recognition research efforts are made to minimize their salience by selecting speakers with similar accents and by having speakers read controlled passages rather than engage in spontaneous speech. Organic characteristics include voice qualities such as fundamental frequency, resonances, nasality, hoarseness, etc. Even in the absence of obvious speech habits, these characteristic qualities of individual voices permit excellent recognition in everyday situations. Organic voice characteristics may be particularly useful in automatic (or computer) recognition of speakers since they are most likely to provide a consistent and measurable set of voice parameters. Nolan (1983, p. 27-28) points out that the organic-learned distinction is simplistic and that voice characteristics may overlap in their classification. It is true that a variety of factors contribute to the actual speech output. For example, speakers may to some extent vary their accent and choice of vocabulary depending on who they are talking to, and organic characteristics such as pitch are in some part under voluntary control and may also vary with emotional state. Still, the organic-learned dichotomy may serve a useful conceptual purpose.

Recognition research with listeners has included both verification and identification tasks as well as known and unknown speakers. Speaker selection, test method, and choice of speech materials are among the factors that must be considered in the measurement of speaker recognition. Speaker selection is by far the most complex problem and has the most far-reaching

consequences for the experimental results. The diversity of experimental methods, which is to be expected in view of the different motivations for different experiments, often makes it difficult to compare results between experiments. The content and duration of the voice sample will also influence recognition.

The ability of listeners to recognize voices varies considerably with the size and composition of the speaker set. Increasing the size of the speaker set increases the difficulty of recognition (Pollack, Pickett, and Sumbly, 1954). Stevens et. al. (1968) found that there were differences in recognition rate with different choices of speaker sets. In general, the recognition rate can be expected to decrease if the speakers are selected to be more homogeneous. Homogeneity of the speaker set is difficult to define as it depends on characteristics that are not always easy to quantify or even identify. Homogeneity has been controlled principally by selecting speakers of the same sex and a similar linguistic background. Other voice characteristics (physical measures such as fundamental frequency or rated voice qualities such as hoarseness) could be added to these in order to select highly similar speakers, and carried to an extreme, this would obviously make the listener's task very difficult. The degree to which it is desirable to control speaker homogeneity depends on the purpose of the research. Differences in speaker selection make it difficult to compare directly the results obtained by different investigators, and for comparison purposes it would be desirable to use a standard speaker set or to develop clearly defined selection criteria. For other purposes, such as investigating how listeners behave in a naturalistic environment, the selection of a large and varied speaker set is likely to be more representative.

The use of known or unknown speakers will largely determine the test method. For a set of known speakers, a naming task would be most appropriate, whereas for an unknown speaker set matching-to-sample or a same-different task would often be a better choice. The effect of speaker set size can be expected to be greater for unknown than for known speakers as a larger speaker set increases the memory load for any memory dependent tasks, and the less memory dependent test methods such as paired comparisons or matching-to-sample (if it requires repeated sample presentation) rapidly become too cumbersome or time-consuming as the size of the speaker set is increased beyond a relatively small number.

The content (vowels, words, sentences) and duration of the voice sample also contribute to the proportion of correct identifications (Bricker & Pruzansky, 1966; Pollack, Pickett, and Sumbly, 1954). Performance increases with the duration of the sample and improves from vowels to monosyllables to disyllables to sentences. However, with a reasonably varied phonemic content, a sample duration of 2-3 seconds seems to be quite sufficient for identification performance to reach an asymptote. Generally the speech samples used in recognition experiments tend to be selected and controlled rather than spontaneous utterances. Some verification procedures may require the use of set phrases, but most occasions for recognizing voices in real-life situations are likely to be conversational utterances, which are quite different from read or rehearsed speech.

Numerous studies have found that some speakers are correctly identified more often than others and that this can depend on the composition of the

speaker set (e.g. McGehee, 1937, 1944; Stevens, et. al., 1968); but there seems to be little research relating specific speaker characteristics directly to identifiability. Certain characteristics such as sex of the speaker or an obvious regional dialect may be easily identified, whereas other qualities that contribute to voice recognizability may be more difficult to characterize and attempt to quantify. Several researchers have used rating scales (often combined with factor analysis or multi-dimensional scaling) to characterize large sets of voice samples (Clarke & Becker, 1969; Holmgren, 1967; Singh and Murry, 1978; Voiers, 1964; Voiers, 1979). Such ratings can either be used descriptively or combined with decision rules to discriminate among speakers. Clarke & Becker (1969) found that identification performance by listeners is superior to that achieved by the use of rating scales. In everyday life, of course, the familiarity of the speaker and other characteristics that contribute to voice distinctiveness will be badly confounded.

I. EXPERIMENT I

The first experiment was concerned with the ability to recognize the voices of known speakers over a voice communication link (LPC) that was known to be degraded in quality. The identification of familiar speakers presents a variety of problems in the selection of test subjects. It can be difficult to find a sufficient number of people whose voices are known to one another, and some of these people will know one another better than others. In this case, we selected a group of coworkers and asked every participant to rate the familiarity of each coworker prior to the listening tests. This procedure also permitted the analysis of correct identifications as a function of familiarity. The speakers were also rated as to the distinctiveness of their

voices to determine if this was a factor in voice recognizability. The speaking task was selected to induce the speakers to communicate in a natural manner while avoiding such complete freedom of expression that vocabulary, idiosyncratic expressions, and sentence structure would serve as the dominant identifying factors. Reading a standard passage gives each speaker exactly the same set of voice materials, but people do not usually read the way they talk. The abbreviated battleship game used in the NRL Communicability Test (Schmidt-Nielsen and Everett, 1982) employs a limited vocabulary in a constrained format for exchanging information, and this task permits the speakers to use a natural, conversational manner and at the same time guarantees a basically similar vocabulary and format across the different recording sessions.

A. Method

The experimental population was a branch consisting of a little over 40 people. Of these 24 served as speakers and an overlapping but not identical group of 24 served as listeners. Most of these people had been in the same branch for several years and for the most part knew each other quite well. Several months elapsed from the beginning of the recording phase to the end of the testing phase, and there were some departures and a few new arrivals during this time. Familiar people who had left were included among the speakers, but new people were not added to the test population. In the recognition phase of the experiment, the entire branch was considered as the population from which the voice samples could have been taken even though not everyone had actually been recorded.

Voice samples were collected from 24 speakers, 15 males and 9 females. In order to obtain speech samples from people actually talking to one another in a natural manner rather than reading from a text, the battleship game from the NRL Communicability Test (Schmidt-Nielsen and Everett, 1982) was used as the communication task. The limited vocabulary associated with the game assures a reasonable consistency from test to test while the game format provides the motivation for the participants to communicate in a reasonably natural manner. Two people at a time play the game. Each person places two "ships" in a 5 X 5 grid, players take turns shooting at one another by specifying cells in the grid (e.g., bravo three, alfa four, etc.). The game continues until one player has sunk both of the other person's ships. The evaluation questionnaire that normally follows this task was omitted.

Each player was seated in a separate sound booth equipped with a push-to-talk handset containing a Roanwell Confidencer Model 240-100002-653 dynamic microphone. The experimenter's control station communicated with both test stations and allowed the experimenter to give instructions and to monitor the sessions, which were tape recorded. The 24 speakers were recorded two at a time. Each pair of speakers completed two battleship games, the first over an unprocessed channel and the second over an LPC voice processing system using the standard DoD LPC-10 algorithm (Tremain, 1982). These LPC recordings were later found to be flawed, and the 24 speakers were all rerecorded (with different playing partners) playing a single game over the LPC system. The second recording was used for the LPC test tapes. Some of the speakers had had extensive experience in talking over LPC systems, but for the majority it was their first exposure to an LPC voice processor. Most of the speakers had little or no difficulty performing the task. Even though a few people

experienced some initial problems understanding one another, all managed to complete the game.

The unprocessed and LPC recordings were dubbed and manually spliced to separate the utterances of the two speakers and to remove the intervening silences. Certain individually distinctive words or phrases and extraneous sounds such as laughs were also deleted. For example, if one player called the other by name, the name was removed, and the expression "You turkey," used by one of the speakers, was deleted. Ordinary variations on the phrases used in playing the game were left as they had been spoken. Thus "My shot is alfa three," "I shoot alfa three," or just "Alfa three" were all left intact, and responses might include "You got me," "Hit," or "Alfa three is a hit." This editing produced two sequences of shots and responses for each speaker, one unprocessed and one LPC sequence. The sequences varied in length depending on the number of moves needed to complete the game as well as on individual talking style. All sequences contained at least 12 utterances and were more than long enough to identify any speaker that could be recognized. (The shortest sequence was 15.4 s and was correctly identified by 100% of the listeners.) The average duration of the unprocessed sequences was 29.8 s with a standard deviation of 9.4 s. For the LPC sequences, the average duration was 54.1 s with a standard deviation of 27.0 s.

Sequences similar to those of the 24 speakers from the branch were also made from tapes of four outside speakers who had participated in a previous experiment and were unknown to the branch members. This group was included to check for guessing strategies.

Two different randomizations of all of the speaker sequences were assembled for the unprocessed tapes and two for the LPC tapes. The number of the sample was recorded before each speaker sequence by a female speaker who was not a member of the speaker set, and the samples were separated by 5 s of silence.

The listeners were recruited from the same branch as the speakers. Of the 24 people who participated in the identification tests, 19 had also been speakers and 5 had not. Before identification testing began, rating questionnaires were handed out to branch personnel asking them to rate their familiarity with each person's voice and the distinctiveness of the voice for each of the 39 people listed on the questionnaire. The ratings were made using 7-point scales from "Totally Unfamiliar" to "Very Familiar" and from "Not Distinctive" to "Highly Distinctive." On the distinctiveness scale, raters could mark a separate box if they felt a person's voice was too unfamiliar to rate accurately.

An Ampex tape recorder equipped with KOSS Pro 4AA headphones was set up in a quiet conference room, and the listeners were tested on one tape at a time whenever a convenient time could be found. Half the listeners heard an LPC tape first and half heard an unprocessed tape first. Each of the randomizations of the LPC tapes and the unprocessed tapes was heard by half the listeners.

Each listener was instructed in the use of the tape recorder and was given a numbered answer sheet on which to fill in the name of the speaker next to the number corresponding to the announced sample number on the tape. The

listeners were allowed to listen to a sample as often as necessary in order to identify the speaker, but they were told not to go back and listen to any previous samples after they had continued to the next sample. A list of 39 branch members (29 M, 10 F) was available for reference, but the listeners were carefully instructed not to check off names or to use any process of elimination because a speaker might be sampled more than once (there actually were two samples for one of the speakers.) The listeners apparently obeyed these instructions since many of them did list several people twice, and four new branch members who were not on the reference list appeared as responses anyway. No mention was made that there might be speakers from outside the branch on the tapes. People were allowed to leave a space blank if they could not identify the speaker (this might happen if the person was unknown to them) but were encouraged to fill in as many as possible. In addition to naming each speaker, the listeners were asked to evaluate their confidence in their answers on a 3-point scale with the labels "guessing," "fairly sure," and "very sure."

B. Results

1. General results

Percent correct identification for each speaker was based on the number of listeners (out of the total 24) who correctly identified the speaker. For the 19 speakers who also served as listeners, the self identifications were excluded, and scores were based on the other 23 listeners. Self identifications will be discussed separately below. Separate scores were computed excluding the responses of listeners who gave a rating of "totally unfamiliar" to a particular speaker. The average percent correct

identifications for the unprocessed samples and for the LPC samples is shown in Table 1. As might be expected, the difference between unprocessed and LPC speech was significant¹ $t(23) = 7.15$. Although speaker recognition dropped for speech over LPC (from nearly 90% correct for unprocessed speech to around 70% correct), the recognition rate for familiar speakers was still very high for LPC speech. This rate is probably fairly representative of situations where a limited vocabulary is required and can be expected to be even higher in informal conversations where more of the individual speaker's speech habits are present as cues for identification.

There were large individual differences in identification rate both for individual speakers and among listeners. This is consistent with the findings of other investigators (e.g. Bricker & Pruzansky, 1966; Stevens et. al., 1968; Rosenberg, 1973). The percentage of the time a speaker was correctly identified ranged from 21.7% to 100% for LPC samples, with a standard deviation of 23.2, and from 47.8% to 100% for unprocessed samples, with standard deviation of 14.8. All speakers who were identified well over LPC were also identified well unprocessed, but several speakers with high unprocessed scores were poorly identified over LPC. Overall, the correlation between speakers' identification scores for unprocessed and LPC speech was 0.66, which is statistically significant but not overwhelmingly high. Listener accuracy ranged from 39.1% to 95.7% for LPC, with a standard deviation of 17.6, and from 64% to 100% for unprocessed with a standard deviation of 9.5. The finding of greater variability among speakers than among listeners is consistent with results found in intelligibility testing (Voiers, 1981). This underlines the importance of using relatively large

¹All tests of statistical significance were conducted using $p < 0.01$. Unless otherwise stated, this value can be assumed where significance is reported. Values of the computed statistic are given for those who are interested in a more exact p value.

Table I. Identification responses for LPC and unprocessed speech samples.

	LPC		with "unfamiliar"		without "unfamiliar"		with "unfamiliar"		Unprocessed		without "unfamiliar"	
	%	n	%	n	%	n	%	n	%	n	%	n
Correct identifications	68.6	382			70.6	374	88.2	491			90.4	479
Don't know	14.7	82			13.6	72	7.0	39			5.3	28
Incorrect identifications	16.7	93			15.8	84	4.8	27			4.3	23
Total	100.0	557			100.0	530	100.0	557			100.0	530

speaker sets to insure a reasonable generality of the results in experiments of this type.

2. Identification and familiarity.

The numbers 0 to 6 were assigned to familiarity rating categories with 0 representing "totally unfamiliar" and 6 representing "highly familiar". Since the listeners and speakers were all selected from a branch of coworkers, most of the ratings were in the high familiarity categories. Because of the small number of 1 and 2 ratings, these two categories were combined. There were a total of 576 responses for each part of this experiment (24 listeners x 24 speakers). The distribution of these responses for each of the rating scales can be seen in Table II.

The way people rated their own voices was highly variable. Some people rated themselves as very familiar and others as totally unfamiliar. Because of this variability, these responses were all combined in a single "self" category. It seems likely that those who rated themselves as familiar felt that they knew their own voice very well because they hear it so often, whereas those who rated themselves as unfamiliar may have realized that one doesn't hear one's own voice as others do.

Of the 19 cases in which a listener's own voice was among the voices to be identified, 16 (84.2%) correctly identified their own unprocessed voices, and 13 (68.4%) correctly recognized their LPC voices. These percentages are very similar to the overall recognition rate for the entire set of voices. One's own voice can be viewed as both familiar and unfamiliar in that it is heard frequently but not in the same way as others hear it. In this group of coworkers who interact with one another regularly, people recognized their own voices about as well as they recognized each other's voices.

Table II. Number of responses in each rating category for the familiarity and distinctiveness ratings.

Familiarity		Distinctiveness	
Category	n	Category	n
0 (totally unfamiliar)	27	Don't know well enough	38
1-2	34	0-2	66
3	40	3	90
4	71	4	133
5	124	5	97
6	261	6 (highly distinctive)	133
Self	<u>19</u>	Self	<u>19</u>
Total	576	Total	576

The percent of correct identifications for all responses in each familiarity category is shown in Figure 1. The proportion of responses for each of the three confidence ratings is also indicated. For both LPC and unprocessed samples, both correct responses and rated confidence in the identifications increased with increasing familiarity of the speakers. Contingency chi-square tests were conducted using 5 levels of identification (incorrect responses, don't knows, and the 3 confidence levels for correct responses) and 5 levels of familiarity (1-2, 3, 4, 5, 6). The zero familiarity responses were not included in this analysis because in theory it should not be possible to identify unknown speakers. The difference in response pattern across familiarity ratings was significant for both LPC, $\chi^2(16) = 32.04$, and unprocessed, $\chi^2(16) = 57.85$, speech samples.

Of the 576 individual ratings, there were 27 instances in which a listener had rated a particular speaker as "totally unfamiliar." However, the putatively unknown speaker was correctly identified by the listeners 44% of the time for the unprocessed samples and 30% for the LPC samples. With a list of about 40 people to choose from, these scores are well above chance levels (even taking into account that males and females were almost never confused). This could mean either that a voice was better known than the rater claimed or that a successful guessing strategy was used. The accuracy of the familiarity ratings could not be checked. However, two possible guessing strategies were considered: an elimination strategy in which a listener guesses people he has not heard prior to the sample in question and a least known strategy in which he limits guesses to the names that are the least familiar. If an elimination strategy were used, most of the errors for the unknown speakers should be from among those branch members who were not in the speaker set, and for the least

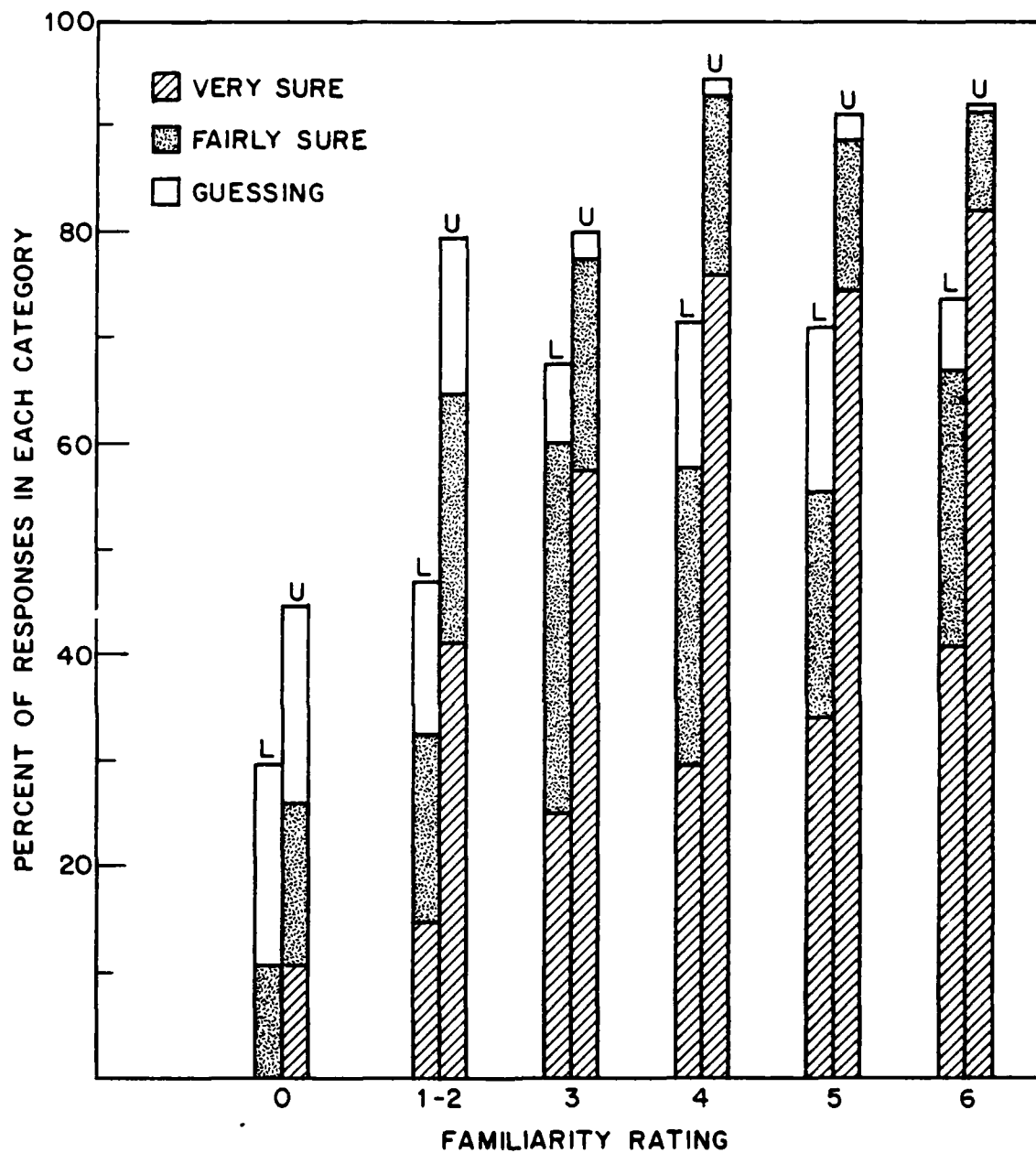


Figure 1. Confidence levels for correct responses as a function of familiarity rating. L--LPC processed speech samples; U--unprocessed speech samples.

known strategy most errors should be from speakers rated low in familiarity. Of the 15 incorrect responses to "unknown" speakers for unprocessed speech, 11 were "don't know" (i.e. the spaces were left blank), 3 were little known branch members who were not speakers, and 1 was a little known speaker. Of the 19 incorrect responses for LPC processed speech, 10 were "don't know", 5 were well known speakers, 2 were little known speakers, and 2 were little known non-speakers. The inclusion of a number of people from the speaker set suggests that an elimination strategy was not used, but guessing from among little known people may have occurred.

An examination of the errors for those speakers rated as familiar revealed primarily two types of errors: inconsistent errors, in which the errors for a given speaker were distributed across several people, and consistent errors, in which one speaker was nearly always mistaken for the same other speaker. The inconsistent errors tended to occur with little known speakers and non-speakers, whereas most of the consistent errors occurred with well known speakers and non-speakers. Confidence ratings indicated slightly more certainty for consistent than for inconsistent errors. Taken together, the error patterns for both familiar and unfamiliar speakers suggest two things: that some errors are confusions due to one person's sounding like another; and that when guessing occurs a least known strategy seems more plausible than an elimination strategy. The fact that several speakers were listed twice by many of the listeners also argues against the elimination strategy. On the whole it seems that if a voice can not be clearly identified as someone who is known, guesses tend to be selected from the least familiar names in the pool of possibilities.

The responses to the four completely unknown outside speakers also seem to follow either a pattern of consistent confusion with a known speaker or a guessing pattern with inconsistent responses. Each of the two unknown females was frequently identified as a single one of the well known speakers, whereas the identifications for the two males were distributed across a number of different speakers. There were also fewer "don't know" responses to the voice of a speaker who was consistently mistaken for one particular individual, than there were when the errors were inconsistent.

In order to test the relationship between guessing strategies and consistent versus variable errors, both kinds of errors were grouped according to the familiarity of each of the wrong responses. Only those speakers for whom there were 3 or more errors were considered. Responses were classified as consistent if at least half of the wrong guesses for a given speaker were the same person, and all other wrong guesses were classified as variable errors. In order to have a sufficient number of responses in each category, the familiarity ratings were combined into 4 categories (0-2, 3-4, 5, and 6). Chi-square tests revealed significant differences in the distribution of familiarity ratings for the two types of errors. The consistent errors were more often familiar people, and the variable errors were more often unfamiliar people. This pattern held both for speakers who belonged in the branch, $\chi^2(3) = 21.37$, and for the four truly unknown speakers, $\chi^2(3) = 12.92$.

3. Identification and distinctiveness.

The relationship between distinctiveness and correct identifications (with confidence ratings indicated) is shown in Figure 2. It should be noted that this analysis uses the same identification data as the analysis by familiarity

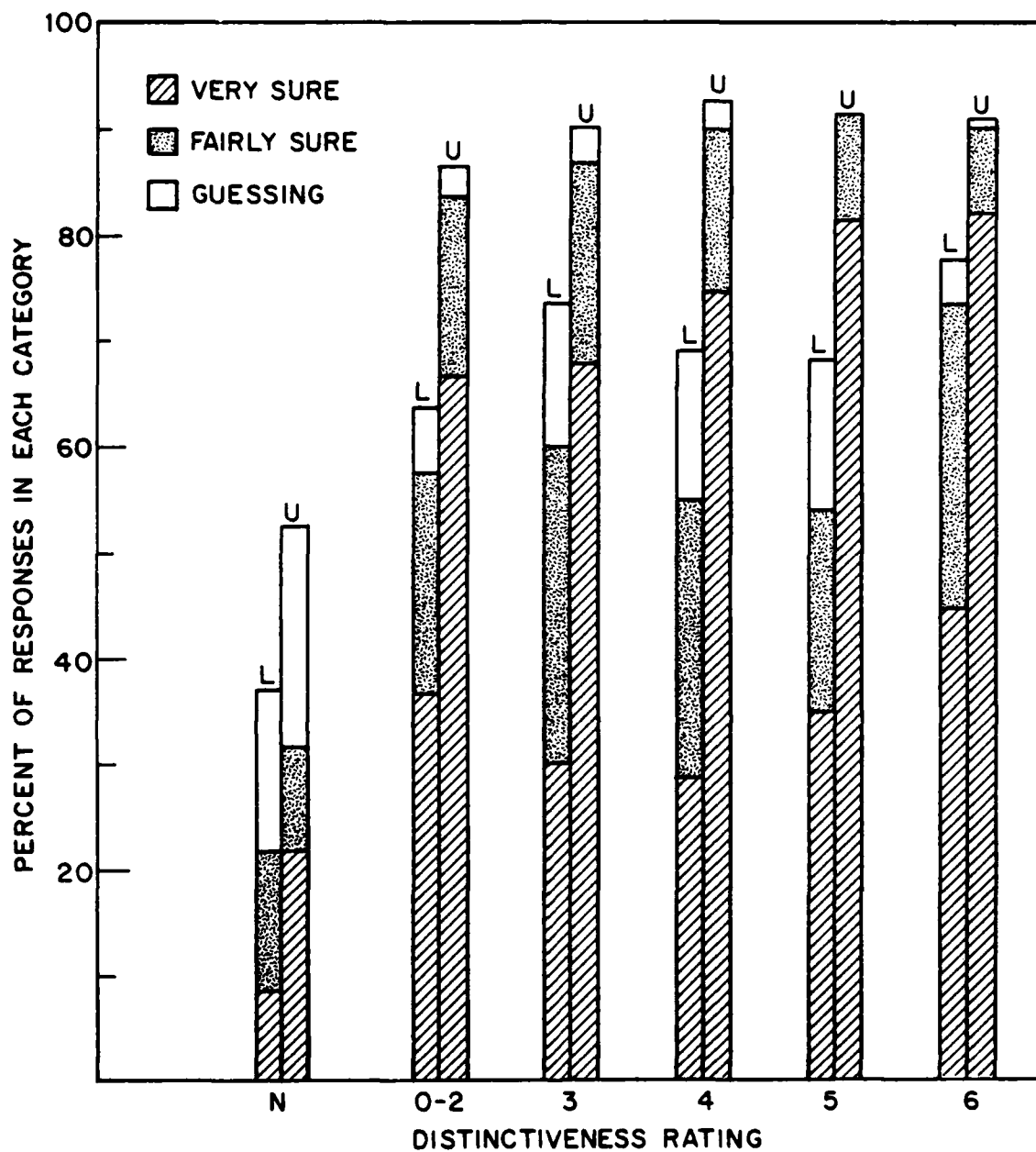


Figure 2. Confidence levels for correct responses as a function of distinctiveness rating. L--LPC processed speech samples; U--unprocessed speech samples.

but with the responses sorted according to distinctiveness ratings; it does not represent an independent set of data, so similarities in the results should be viewed with caution. A Chi Square test showed no significant difference in the pattern of responses across distinctiveness categories for either LPC speech, $\chi^2(16) = 25.60$, or for unprocessed speech, $\chi^2(16) = 23.81$. The scores for the voices not known well enough to give a distinctiveness rating were excluded from this analysis (this group included the voices rated zero on the familiarity scale as well as a few others with very low familiarity ratings). As was true for the familiarity ratings, these voices were identified at a considerably better than chance level. The correlation between average rated distinctiveness for each speaker and the percent of times that speaker was correctly identified was 0.40 for LPC and also 0.40 for unprocessed speech, which does not reach statistical significance.

There was, however, a significant correlation, $r = 0.79$, between rated familiarity and rated distinctiveness. It seems probable that in making their ratings, people tended to feel that a voice was distinctive if it was well known to them. It seems intuitively reasonable that well-known voices should be easier to recognize and therefore might be regarded as being more distinctive. If there is an abstract quality of "distinctiveness" that makes some people's voices more recognizable than others, it was not captured in these ratings made by people who were familiar with the speakers. Both the familiarity and distinctiveness ratings were based on memory, not on the recorded voices, since the rating questionnaires were filled out before the identification tests were given. This insured that the ratings would not be influenced by the subsequent ease or difficulty in recognizing the voices.

The proposition that some voices are more unusual or distinctive than others does not seem an unreasonable one, although the distinctiveness will certainly also be affected by the context of the other voices in the sample. For example, a strong regional or foreign accent can be highly distinctive when it is the only one in an otherwise homogeneous group, and in the present sample the one speaker with a foreign accent was also the only one to be correctly identified 100% of the time for the LPC speech samples. Two males from the same Southern state were sometimes confused with one another, and one of the two female impostors was consistently mistaken for one of the speakers in the sample with a similar accent. There were almost no confusions between male and female speakers, and the one or two that did occur were for the LPC speech, where pitch is not always accurately conveyed.

II. EXPERIMENT II

In the first experiment distinctiveness ratings by people familiar with the speakers did not predict how well the speakers would be recognized. Yet it is apparent from the results of this and other experiments that some people are more often correctly recognized than others, and that this can be true whether or not they are also familiar. However, some characteristics may be distinctive relative to the reference group (e.g., the only foreign voice in a group of native English speakers), and others may be more or less unique to the individual (e.g., a particularly raspy voice or unusual speech habits). The second experiment was conducted to determine whether ratings uncontaminated by familiarity with the speaker would be better at predicting correct identifications.

Voiers (1964,1979) used a large number of rating scales with 200 male voices and applied factor analysis to identify 8 factors for evaluating speaker recognizability. Singh and Murry (1978) used multi-dimensional scaling to characterize a set of 20 voices, 10 male and 10 female, but no actual tests of identification were reported. Clarke and Becker (1969) compared the use of rating scales to discriminate among speakers with other methods for identifying speakers. They developed a set of rating scales for one set of 20 speakers and attempted to discriminate among a second set of 20 speakers by having listeners use the scales to rate the speakers and then applying a set of decision rules. The maximum discrimination using the rating scales was 51%, and both tests with human listeners as well as the use of a limited set of physical measures performed better than ratings in discriminating among speakers. None of these studies attempted to use the rated characteristics to predict how well a speaker would be recognized by listeners. In addition to rating distinctiveness, the subjects in Experiment II also rated the voices on rating scales for five of the eight factors from Voiers' (1979) factor analysis.

A. Method

Volunteers unfamiliar with any of the speakers were recruited through the University of Maryland Psychology Department. There were a total of 54 listeners of whom 27 listened to and rated the unprocessed samples and 27 rated the LPC processed samples. The tapes of spliced sequences from the identification experiment were used to obtain the ratings. The ratings of the unprocessed samples can be compared with the ratings by familiar raters, and the ratings of the LPC samples may give some indication of how voice distinctiveness is altered by LPC processing.

The rating forms consisted of 24 numbered sets of six rating scales. Each scale consisted of a question and a 7-point rating scale with the endpoints labeled. The first question was the one asked of the familiar raters: "How DISTINCTIVE or characteristic do you consider this person's voice?" with the endpoints HIGHLY DISTINCTIVE and NOT DISTINCTIVE. The other scales were the first five factors from Voiers' (1979) factor analysis of rated voice characteristics. These were the factors labeled by Voiers as animation, pitch, continuity, charisma, and roughness. Because of testing time and subject considerations, it was not feasible to administer all of the 75 rating scales that Voiers used. Each of the five factors was presented in the form: "How would you rate the _____ (name of factor) of this voice?" The endpoints for each rating scale were labels from the two rating scales from Voiers' set of 75 scales that were most highly correlated with that factor--i.e., for ANIMATION the scale was from HURRIED, BUSY to UNHURRIED, IDLE. It seems reasonable to expect that this strategy might yield ratings that would be fairly similar to the factors in Voiers' analysis for the purpose of determining if the rated voice characteristics could be related to voice identification rate.

B. Results

The 27 listener ratings on each scale were combined to obtain average ratings for each speaker. Correlations were obtained using these average ratings and the correct identification rates from the first experiment for each speaker. There were two identification scores (unprocessed and LPC) and 12 ratings (6 unprocessed and 6 LPC) for each speaker. Therefore, the same identification scores were repeatedly used in different correlations with each

of the 6 rating scores. In order to hold the overall error rate at $p \leq .05$, a correlation of 0.522 or greater is required when 6 tests are conducted at the same time. The correlations between ratings and correct identification rate are shown in Table III. As can be seen, none of the correlations of ratings with identification rate reaches statistical significance. Since the same tapes that had been used for the identifications were also used for the ratings, this should have increased the likelihood of finding a relationship if one existed. In fact, these data have been so over-analyzed that any trends that might be found should be retested with an independent set of speech samples. It might be argued that the relationship between the rating scales and voice identifiability is not likely to be linear in any case. A more reasonable supposition might be that extreme values (high or low) on any of the scales would make a voice more recognizable. This could be tested by postulating a curvilinear relationship between ratings and correct identifications. Actually a simple test of the hypothesis that voices with extreme ratings would be better identified was carried out, instead of conducting a second regression analysis which would require stronger assumptions about the shape of the function. The six people with the most extreme ratings were selected for each scale, and these groups were compared for overlap with the set of six best identified speakers. For none of the scales was this overlap greater than might be expected by chance; the chance expectation is an overlap of 1.5, and actual overlaps were 1 for four of the scales, 0 for one, and 2 for one. Even the most optimistic scrutiny of the data fails to show any form of consistent relationship between rated voice characteristics and the ability of listeners to recognize known speakers either unprocessed or over an LPC voice processor.

Table III. Correlations between rated characteristics by unfamiliar listeners and correct identifications by familiar listeners.

Ratings of with	Unprocessed speech		LPC speech
	Unproc. % correct	LPC % correct	LPC % correct
Distinctiveness	0.01	0.09	0.45
Animation	-0.46	-0.31	-0.05
Pitch	0.21	0.43	0.42
Continuity	-0.18	-0.31	-0.24
Charisma	-0.13	-0.25	-0.46
Roughness	-0.23	-0.35	-0.43
Unfamiliar distinctiveness		Unprocessed	LPC
Familiar distinctiveness		0.40	-0.14

III. SUMMARY AND CONCLUSIONS

This research was concerned with the recognition of familiar speakers talking in a conversational manner over an unprocessed voice channel and over a degraded voice channel, a 2.4 kbps LPC voice processing system. Speaker recognition over LPC was significantly poorer than with unprocessed speech. Still speaker recognition over LPC was very good considering that for the most part the listeners had never heard these speakers over an LPC system. (Each of the listeners who had also been a speaker had talked with one other person, and three or four of the listeners had previous experience with LPC systems and had talked with several other branch members over such systems.) In spite of the degradation in voice quality brought about by LPC processing, a large part of the information used by listeners to recognize familiar speakers seems to be preserved, at least for conversational speech. In contrast, Shearme and Holmes (1959) found very poor recognition of voices that were shifted in pitch using a vocoder, but the listening tests were paired comparisons with the untransformed speech as the standard. The procedural differences as well as the pitch shift probably account for this result.

Not surprisingly, rated familiarity was significantly correlated with correct identification rate. Rated voice distinctiveness, on the other hand, was not related to correct identification rate. Neither were ratings of distinctiveness and of five other voice characteristics made by listeners unfamiliar with the speakers. The lack of any correspondence between rated voice characteristics and identification rate suggests that the cues used by listeners to recognize familiar voices are not the same as those that were tapped when listeners used the rating scales. It may be that when listeners

perform the identification task, they use cues which they are incapable of bringing to conscious awareness. It is possible that rated voice qualities may be better at predicting the recognition of unknown speakers, and this possibility is currently being investigated. Another possibility is that ratings (at least the ones used here) are not really very good for characterizing voices. Certainly there was a great deal of variability in the voice ratings, both by the familiar and by the unfamiliar raters. The complete lack of correlation between the distinctiveness ratings made by the familiar raters and those made by the unfamiliar raters tends to support the supposition that the rating process itself may be at fault.

Preliminary results of research currently in progress indicate that identifications made by naive listeners to whom the speakers were unknown are significantly correlated with the identifications by the known listeners. This suggests that while the intuitive notion that some voices are more "distinctive" than others is not entirely unjustified, ratings of distinctiveness do not seem to capture this notion. This is somewhat surprising because ratings have been used with considerable success in numerous other applications. For example, rated speech intelligibility is highly correlated with scores on intelligibility tests (Voiers, 1981).

Other investigators have successfully used both rated voice characteristics and selected physical measures to discriminate among voices (e. g., Holmgren, 1967; Clarke and Becker, 1969), although Clarke and Becker found that performance by listeners was better than either of these. Physical voice measures form the basis of automatic speaker recognition systems, and Rosenberg (1973) found that an automatic system could perform a

customer-imposter verification task better than human listeners. In this study no physical measures of voice characteristics were made, and it is possible that a relationship with listener recognition rate might be found using such measures.

The results of research investigating the ability of listeners to recognize familiar voices were reported. Identification using high quality tape recordings was compared with performance when the speakers were talking over a narrowband LPC voice communication system. Identification of the most familiar listeners was 94% unprocessed and 74% over LPC; and identification of the least familiar listeners was 44% unprocessed and 30% over LPC. Rated characteristics of the speakers' voices were not found to be related to identification rate. Further research is needed to determine if other voice characteristics can be better related to identification rate by listening.

REFERENCES

- Atal, B. S. (1976). "Automatic recognition of speakers from their voices," Proceedings of the IEEE. 64, 460-475.
- Bricker, P. D. and Pruzansky, S. (1966). "Effects of stimulus content and duration on talker identification," J. Acoust. Soc. Am. 40, 1441-1449.
- Bricker, P. D. and Pruzansky, S. (1975). "Speaker recognition," in: Lass, N. J. (ed.) Contemporary Issues in Experimental Phonetics. Springfield, IL: Charles C. Thomas.
- Clarke, F. R. and Becker, R. W. (1969). "Comparison of techniques for discriminating among talkers," J. Speech Hear. Res. 12, 747-761.
- Garvin, P. and Ladefoged, P. (1963). "Speaker identification and message identification in speech recognition," Phonetica. 9, 193-199.
- Hecker, M. H. L. (1971). "Speaker recognition: an interpretive survey of the literature," ASHA Monograph 16, American Speech and Hearing Association, Washington, DC.
- Holmgren, G. L. (1967). "Physical and psychological correlates of speaker recognition," J. Speech Hear. Res. 10, 57-66.
- McGehee, F. (1937). "The reliability of the identification of human voice," J. Gen. Psychol. 17, 249-271.
- McGehee, F. (1944). "An experimental study in voice recognition," J. Gen. Psychol. 31, 53-65.

- Nolan, F. (1983). The phonetic bases of speaker recognition. New York: Cambridge University Press.
- Pollack, I., Pickett, J. M., and Sumby, W. H., (1954) "On the identification of speakers by voice," J. Acoust. Soc. Am. 26, 403-406.
- Rosenberg, A. E. (1973) "Listener performance in speaker verification tasks," IEEE Trans. Audio Electroacoust. AU-21, 221-225.
- Rosenberg, A. E. (1976). "Automatic speaker verification: A review," Proc. IEEE. 64, 475-487.
- Sandy, G. F. (1982). "Digital Voice Processor Consortium Interim Report," Appendix A, MTR-81W0159-02, Mitre Corp., McLean, VA.
- Schmidt-Nielsen, A. and Everett, S. S. (1982). "A conversational test for comparing voice systems using working two-way communication links," IEEE Trans. Acoust., Speech, Signal Processing. ASSP-30, 853-863.
- Shearme, J. N. and Holmes, J. N. (1959) "An experiment concerning the recognition of voices," Lang. Speech. 2, 123-128.
- Singh, S. and Murry, T. (1978) "Multidimensional classification of normal voice qualities," J. Acoust. Soc. Am. 64, 81-87.

Stevens, K. N., Williams, C. E., Carbonell, J. R., and Woods, B. (1968) "Speaker authentication and identification: A comparison of spectrographic and auditory presentations of speech material," J. Acoust. Soc. Am. 44, 1596-1607.

Tremain, T. (1982). "The government standard linear predictive coding algorithm: LPC-10," Speech Tech. 1 (2).

Voiers, W. D. (1964) "Perceptual bases of speaker identity," J. Acoust. Soc. Am. 36, 1065-1073.

Voiers, W. D. (1979). "Toward the development of practical methods of evaluating speaker recognizability." ICASSP 79. 1979 Int. Conf. Acoust. Speech Signal Process., Washington, DC, 2-4 Apr., 1979. 793-796 (IEEE, New York, 1979).

Voiers, W. D. (1981) "Uses, limitations, and interrelations of present-day intelligibility tests," presented at the National Electronics Conference, Chicago, Oct., 1981. Vol. 35, pg. 251.